DOCUMENT RESUME

ED 128 375 TM 005 497

AUTHOR Garwick, Geoffrey
TITLE Program Evaluation Project Report, 1969-73. Chapter Three: An Introduction to Reliability and the Goal Attainment Scaling Methodology.
INSTITUTION Program Evaluation Resource Center, Minneapolis, Minn.
SPONS AGENCY National Inst. of Mental Health (DHEW), Rockville, Md. Div. of Mental Health Services Program.
PUB DATE Jun 74
GRANT NIMH-5-R01-1678904
NOTE 19p.; For related documents, See TM 005 495-501
AVAILABLE FROM Program Evaluation Project, 501 Park Ave. South, Minneapolis, Minnesota 55415 ($1.00)

EDRS PRICE MF-$0.83 HC-$1.67 Plus Postage.
DESCRIPTORS *Evaluation Methods; *Goal Orientation; Interviews; Measurement Techniques; *Mental Health Programs; *Program Evaluation; *Reliability; Scores
IDENTIFIERS *Goal Attainment Scaling

ABSTRACT

The P.E.P. Report 1969-1973 focuses on the various findings and activities of the Program Evaluation Project. Reliability is considered a basic aspect of any measurement system. With Goal Attainment Scaling, at least two types of reliability are important: the reliability of the followup guide construction and the reliability of the followup guide scoring. This chapter discusses the theory underlying applications of conventional reliability concepts to Goal Attainment Scaling and reviews a range of studies relevant to the reliability of the methodology. This chapter is designed to give a general introduction to reliability and Goal Attainment Scaling. (Author/RC)

P.E.P. REPORT 1969-1973

CHAPTER THREE

AN INTRODUCTION TO RELIABILITY AND THE GOAL ATTAINMENT SCALING METHODOLOGY

A REPORT ON FOUR YEARS OF STAFF EFFORT AT THE PROGRAM EVALUATION PROJECT.

CHAPTER THREE

Program Evaluation Project Report, 1969-1973

AN INTRODUCTION TO RELIABILITY AND THE GOAL ATTAINMENT SCALING METHODOLOGY

(A Review of the Program Evaluation Project Reliability Studies)

Geoffrey Garwick

June, 1974

Thomas J. Kiresuk, Ph.D., Director
Program Evaluation Project
501 Park Avenue South
Minneapolis, Minnesota 55415

Developed under Grant #5 R01 1678904, National Institute of Mental Health, Department of Health, Education, and Welfare

---

Acknowledgements are given to: Dr. David Weiss, for his many helpful comments. Donna Audette, James Baxter, Carolyn Jasperson, Mary Ellen Whalen and other Program Evaluation Project staff members for assistance in locating and developing portions of the data included here. Drs. Sherman and Kiresuk for discussions on Goal Attainment Scaling theory.

TABLE OF CONTENTS

For further information, please contact Ms. Joan Brintnall, Program Evaluation Project, 501 Park Avenue South, Minneapolis, Minnesota 55415.

GENERAL INTRODUCTION TO THE P.E.P. REPORT 1969-1973

The P.E.P. Report 1969-1973 focuses on the various findings and activities of the Program Evaluation Project. It is being published in pamphlet form with one pamphlet for each chapter.

As of January, 1974, the Program Evaluation Project is funded by a three year collaborative grant with the Mental Health Services Division of the National Institute of Mental Health. The purpose of the grant is to emphasize the coordination and dissemination of information on a variety of program evaluation methodologies. Currently, it is expected that the title of the organization will be changed to the Program Evaluation Resource Center during 1974.

Further information on the Goal Attainment Scaling methodology and program evaluation is available in other written and recorded materials from the Program Evaluation Project office. Chapter One, "Basic Goal Attainment Scaling Procedures"; Chapter Five, "A Construct Validity Overview of Goal Attainment Scaling"; and Chapter Nine, "Evaluation of the Adult Outpatient Program, Hennepin County Mental Health Service" of the P.E.P. Report 1969-1973 are now available. Additional chapters will be released this year as they are completed.

SYNOPSIS FOR CHAPTER THREE

AN INTRODUCTION TO RELIABILITY AND THE GOAL ATTAINMENT SCALING METHODOLOGY

PURPOSE: Reliability is considered to be a basic aspect of any measurement system. With Goal Attainment Scaling, at least two types of reliability are important: the reliability of the follow-up guide construction and the reliability of the follow-up guide scoring. This chapter discusses the theory underlying applications of conventional reliability concepts to Goal Attainment Scaling and reviews a range of studies relevant to the reliability of the methodology. This chapter is designed to give a general introduction to reliability and Goal Attainment Scaling. Another P.E.P. Report 1969-1973 chapter on reliability discusses one particular study in dept and will be released later.

MAJOR FINDINGS: In the examination of Goal Attainment Scaling reliability, it should be remembered that any outcome evaluation methodology is designed to vary from one point of measurement to another. Thus, because Goal Attainment Scaling is such an evaluation-oriented measurement, it should not be expected to produce identical results if the same client is scored at different times. As a result of this characteristic and of difficulties in applying other conventional reliability approaches to Goal Attainment Scaling, most of the available reliability studies concentrate on inter-rater agreement with a few others being concerned with alternate form reliability.

In the original reliability study, which is discussed in greater detail in another P.E.P. Report 1969-1973 chapter, for each of 44 clients at the outpatient unit, one follow-up guide was constructed by the intake interviewer and a second follow-up guide was made somewhat later by the therapist. These two follow-up guides were combined and then scored twice at two separate interviews by two different raters. For the follow-up guide prepared by the intake interviewer, the Goal Attainment scores from the two interviews were correlated .711 and for the follow-up guide prepared by the therapist, scores from the two interviews correlated .625. The other P.E.P. Report 1969-1973 chapter on reliability discusses this study more intensively (Sherman, Baxter, Audette, 1974).

A variety of other reliability studies are discussed in less detail. In the interdisciplinary reliability study, 60 clients were interviewed twice on the basis of follow-up guides constructed by intake interviewers, with the interviews being conducted either by nurses or social workers and either by telephone or in person. For this study, the Goal Attainment scores from the first and second interviews were correlated .65, and there were not significant differences in mean scores between the two types of interviewers or between the telephone and in-person interviews.

Other studies cover data from multiple raters scoring a videotape interview, multiple rater scorings of organizational goals, comparison of client and follow-up interview ratings (correlation of .71), and multiple ratings over time in the drug effectiveness study.

I. Theory and Background on Goal Attainment Scaling Reliability

Reliability is a basic property of an instrument or methodology designed for practical use. (Goal Attainment Scaling is generally referred to in this discussion as a "methodology" or "technique" rather than a "test", since tests are often interpreted as tasks done by an individual rather than tasks which are done for or to the individual. [Kelly, 1967]). Goal Attainment Scaling is a relatively unique methodology, and because of its nonstandardized, outcome-oriented nature, most of the classical concepts of reliability have to be interpreted loosely when applied to it. However, since reliability is of such central importance to many persons utilizing the methodology, the author in this chapter shows one way in which Goal Attainment Scaling reliability could be approached.

This section of the chapter discusses both the theory and data on reliability concepts as applied to Goal Attainment Scaling. The theoretical discussion which begins this section of the chapter is followed by a summary of several reliability investigations undertaken by the Program Evaluation Project staff. (As noted in the synopsis, a second chapter on reliability written by Sherman, Baxter, Audette, presents a more intensive analysis of the original Program Evaluation Project reliability study.)

In this chapter, the psychometric complexity of studying reliability in a new instrument is underscored. According to Cronbach, one of the major psychometric theorists, "any research based on measurement must be concerned with the accuracy or dependability or as we usually call it reliability of measurement. A reliability coefficient demonstrates whether the test designed was correct in expecting a certain collection of items to yield interpretable statements about individual differences." (1951) Although such a definition can at least conceptually be applied to the reliability of Goal Attainment Scaling, there are at least two immediate problems in such an application. These two problems are summarized in the pair of Cronbach's phrases:

- "...a certain collection of items..."

  and

- "...statements about individual differences."

When Cronbach refers to "...a certain collection of items...", he is proceeding from the typical background of testing theory, which is based largely on fixed-item aptitude or information tests and inventories of opinion. However, Goal Attainment Scaling is not a "test" in many of its applications, as noted previously, but an indicator of outcomes. Thus, the Goal Attainment score is not basically a "...statement about individual differences". In fact, Goal Attainment Scaling is designed to minimize the impact of individual differences on the Goal Attainment score by orienting measurement to the "expected" or "best prediction" outcome. Goal Attainment Scaling is usually based on a flexible (as opposed to "certain") set of items since the scales are individually developed for each client.

A. Tests of Reliability Appropriate to Goal Attainment Scaling

An indicator of outcomes of treatment or other processes must be able to vary from one point of measurement to another point (usually these are points in "time" since most treatment proceeds through a time period) if there is to be any possibility of evaluating the treatment. One of the assumptions of current theories of treatment outcome is that most people do change (spontaneously or otherwise) during mental health treatment and that these changes occur at differing rates over time.

A similar situation exists with respect to the Minnesota Multiphasic Personality Inventory, and it is reported that: "The limitations in the methods of reliability estimation based upon retest data can be readily discerned as they apply to instruments like the Minnesota Multiphasic Personality Inventory. The usual index of score stability is based upon the degree of correspondence between the ranking of subjects in a group on two different occasions, summarized in a correlation coefficient. To interpret such a correlation as a gauge of the dependability of the scores on some scales, it is necessary to assume that the rankings of the group from the first to the second testing should not change except through errors in the measurement of their positions. Since there is scarcely any scale on the Minnesota Multiphasic Personality Inventory for which this general assumption is tenable for any period longer than a day or two, the various estimates of scale stability published on the basic Minnesota Multiphasic Personality Inventory scales cannot be readily interpreted as indices of the inherent dependability of the scores from these scales obtained on any one occasion." (Dahlstrom, 1969) A similar observation was made much earlier by Hathaway (1956), who argues that:

"It is always difficult to evaluate any of the usual reliability data on personality measures that are likely to show valid time-related variance in the individual subjects. All the Minnesota Multiphasic Personality Inventory scales are sensitive to therapeutic and other effects. Since both the motivation and the life situation of the subjects are likely to change almost momentarily, it is always possible that an observed change in score is valid variance instead of error variance." (1956) Like the MMPI theorists, the Rorschach expert Holzberg (1960) contends that "it cannot be assumed that the object of study, the personality of a subject, is unchanging. Significant aspects of the person-

ality change through time in response to internal and external factors," (p. 368) and concludes "...that the traditional methods of assessing psychometric reliability are inappropriate to the Rorschach ..." (p. 377).

Anastasi (1968) states that "...in its broadest sense, test reliability indicates that extent to which individual differences in test scores are attributable to true differences in the characteristics under consideration and the extent to which they are attributable to chance errors... Factors that might be considered error variance for one purpose would be classified under true variance for another. For instance, if we are interested in measuring fluctuations of mood, then the day-by-day changes in score on a test of cheerfulness-depression would be relevant to the purpose of the test and would hence be part of the true variance of the scores." Thus, experience with other test systems, all of which share some but not all of the sources of reliability associated with the Goal Attainment Scaling methodology, suggest that variation over time does not always indicate a deficiency in the measurement instrument. In the case of outcome-oriented measurements, change over time could be meaningful variation, rather than "unreliability".

The above citations refer basically to "retest" measures of reliability. There is much disagreement among psychometricians as to the proper classification of the different approaches to reliability estimation. Anastasi, however, advances a useful list of five types of reliability estimates.

a. Test-retest (giving the same test twice). Uses two or more testing sessions.

b. Alternate-form (also called parallel forms by other experts). Uses only one testing session.

c. Split-half (dividing the test into two halves). Uses only one testing session.

d. Kuder-Richardson (the internal consistency or inter-item homogeneity approach). Uses only one testing session.

e. Scorer accuracy (accuracy of scoring the test). Uses only one testing session.

As already emphasized, almost all applications of reliability theory to Goal Attainment Scaling involve some extension or special modifications of theory. The application to the Goal Attainment score of each of these five types of reliability estimates is discussed below.

TYPE 1. TEST-RETEST RELIABILITY

Because the Goal Attainment score would be expected to vary meaningfully over time, reliability measurements should be based on observations which are essentially carried out at the same time. When test-retest scores are used for reliability estimates, any significant gap in time between test and retest will deflate the reliability estimate. Thus, the test-retest approach to reliability estimation is probably not appropriate to Goal Attainment Scaling.

TYPE 2. ALTERNATE FORM RELIABILITY

When one follow-up guide is constructed by the client's intake interviewer and a second follow-up guide is constructed by the client's therapist, as in the original reliability study, it could be said that two "alternate forms" have been developed. Many other pairs of alternate forms could be constructed, of course, such as client-constructed follow-up guides versus therapist-constructed follow-up guides, and so on. Most investigations of Goal Attainment Follow-up Guides' content reliability depend, in effect, on the construction of "alternate forms" of a Goal Attainment Follow-up Guide for a client.

However, the construction of two or more follow-up guides for the same client might more simply be conceptualized as a form of "inter-rater agreement" reliability. If two or more raters have a high degree of agreement in the contents and outcomes as measured by the follow-up guides they construct, the reliability estimate is high. Of course, the usual concept of inter-rater agreement does not involve the actual development of the items by the raters, but in the case of Goal Attainment Scaling, the construction of the scales, especially the expected levels, could be thought of as a form of predictive rating of the client's expected outcome.

TYPE 3. SPLIT-HALF RELIABILITY

This reliability estimate is not appropriate to Goal Attainment Scaling. It involves splitting the test into comparable halves by some method and is a form of internal consistency reliability.

This reliability estimate is based on the concept that variance should be similar for both halves of the instrument and that the scales will have high inter-correlations. Such assumptions do not hold for Goal Attainment Scaling.

TYPE 4. INTERNAL CONSISTENCY RELIABILITY

The same reservations cited for the Split-Half reliability estimate apply even more strongly for the Kuder-Richardson method, which is an extension of the internal consistency concept. This approach is probably not useful where the scores or contents are expected to vary over time.

Nunnally (1967), after presenting the possibility of meaningful (i.e., valid) changes in subjects over time argues that "Systematic differences in contents of tests of variations in

people from one occasion to another cannot be adequately handled by a model which is based on the random sampling of items. For adequately handling these factors, the model must be extended to consider the random sampling of whole tests, in which case the tests are thought of as being randomly sampled for particular occasions and correlations among tests are permitted to be somewhat lower than would be predicted from the correlations among items within tests. In that case, the average correlation among a number of alternative forms administered on different occasions, or the correlation between only two such forms, would be a better estimate than that provided by coefficient alpha or KR-20." Nunnally's arguments imply that for tests measuring variables expected to vary over time, such as Goal Attainment Scaling, internal consistency estimates of reliability are not necessarily appropriate.

The related alpha coefficient of Cronbach also demands inter-item homogeneity. As noted before, homogeneity is not necessarily assumed for the Goal Attainment Follow-up Guide of a client.

TYPE 5. SCORER ACCURACY RELIABILITY

For observations where there is some degree of subjectivity in the scoring or rating, two or more scorers can rate each result. The resulting scores are correlated to produce an estimate of the agreement or accuracy of the scoring. In the case of Goal Attainment Scaling, it could be said that there is a second type of "scorer accuracy" reliability estimate based on the similarity of contents on follow-up guides for two or more raters (as mentioned above, this use of the inter-rater agreement idea in this context is unusual but possible!).

A number of studies of Goal Attainment Scaling reliability have been undertaken, but almost all involve the "inter-rater agreement" approach to reliability measuring inter-rater agreement on either the construction of the Goal Attainment Follow-up Guide or scoring of the Goal Attainment Follow-up Guide. Figure I lists the Program Evaluation Project reliability studies which are discussed in this chapter.

FIGURE I: Program Evaluation Project Studies of Goal Attainment Scaling Reliability

STUDY	INTER-RATER AGREEMENT IN CONSTRUCTION OF GOAL ATTAINMENT FOLLOW-UP GUIDES (SCORER ACCURACY OR ALTERNATE FORM)	INTER-RATER AGREEMENT IN SCORING GOAL ATTAINMENT FOLLOW-UP GUIDES (SCORER ACCURACY)
ORIGINAL RELIABILITY STUDY	Intake Interviewer vs Therapist Goal Attainment Follow-up Guides	Rater 1 vs Rater 2 (Two sessions)
INTERDISCIPLINARY RELIABILITY STUDY		In-person vs Telephone Rater 1 vs Rater 2 (Two sessions)
RE-DESIGN VALIDITY/ RELIABILITY STUDY		Therapist vs Independent Interviewer (Two sessions)
GUIDE TO GOALS STUDY PHASE 2	Client vs Intake Worker Goal Attainment Follow-up Guide Study	Client vs Independent Interviewer (One session)
VIDEOTAPE	Multiple Follow-up Guide Constructors	Multiple Scorers (One session)
DRUG STUDY		Three Raters Time 1 vs Time 2 vs Time 3 (One session per time)

B. Reporting Reliability Results

1. Estimates of Reliability

All single experimental measures of reliability are estimates derived from one particular sample of clients, clinicians, raters, and so on, of the theoretical "true" degree of reliability which is the mean of a theoretical distribution of reliability scores. In practice, except for this hypothetical true mean reliability score, there is no single or absolute reliability correlation for an instrument. The reliability may vary depending on the situation in which the instrument is applied (e.g., if the raters are inexperienced, a lower inter-rater reliability score might be expected. Anastasi notes, for example, that even for the venerable Stanford-Binet, the reliability coefficient varies from .83 to .98 for various ages and I.Q. levels. (Anastasi, 1968)

2. The Coefficient of Correlation and Reliability Estimates

The common use of the coefficient correlation is a matter of tradition or convenience. Reliability estimates are most completely expressed by descriptions of the components of variance which are due to various true score and error components. However, for many persons in the human service field, correlation coefficients are more familiar than analysis of variance of error components. Unless noted otherwise, all coefficients are based on the Pearson Product Moment correlation.

3. Percentage of Agreement and Reliability

At times, various percentage of agreement measures are used. Such measures are intended to suggest the degree of inter-rater agreement of Goal Attainment scores, but are not directly comparable to reliability coefficients.

4. Two Methods of Expressing Goal Attainment Scores

The correlation coefficients may be based on either of two different expressions of the Goal Attainment scores. The first expression is the traditional Kiresuk-Sherman Goal Attainment score, which gives a single, summary score for the entire Goal Attainment Follow-up Guide, that is, one score per client in most cases.

Goal Attainment score:

$$50 + \frac{10 \sum_{i=1}^{n} w_i x_i}{\sqrt{(1-\rho)\sum_{i=1}^{n} w_i^2 + \rho\left(\sum_{i=1}^{n} w_i\right)^2}}$$

The second expression is the scale-by-scale score, which is a simple mean of the scale scores on a single follow-up guide. Each Goal Attainment Follow-up Guide, of course, is made up of several individually developed "scales", which could also be called items. The scale-by-scale scores may be presented either in a -2 or +2 range or a 1 to 9 range. The 1 to 9 range is more convenient for computer entry, so that +2 is equivalent to 9, +1.5 is equivalent to 8, +1.0 is equivalent to 7 and so on.) Scale-by-scale analyses assume, in effect, that the Goal Attainment Follow-up Guide is somewhat akin to an inventory or test composed of a number of semi-Likert-type scales. For various purposes, investigators may be interested in either the total Goal Attainment score or in the scores on individual scales on a follow-up guide.

II. Summaries of Reliability Studies

In this section, the reliability studies will be presented roughly in chronological order, except for the internal consistency measures, which will be presented together in a single subsection. The following comments summarize the correlational results of the original reliability study.

A. The Original Reliability Study Results

The original reliability study was based on two independent scorings of eighty-eight follow-up guides for forty-four cases. These forty-four cases were those in a larger group which met a range of criteria as described in the Sherman, Baxter, and Audette reliability chapter (1974). The study produced the following inter-rater agreement reliability estimates of scorer reliability and follow-up guide constructor reliability. Each client was represented by two follow-up guides, one constructed by the client's intake interviewer and the second constructed by the client's therapist approximately three weeks later. Scales from the two follow-up guides were intermixed into a single composite guide. This composite follow-up guide was scored independently by two different scorers at sessions which averaged 25 days apart. The time elapsed between the two follow-up interviews was unintentional, as it was due to delays in scheduling the second interview. Table I presents the correlations.

TABLE I: Correlation Coefficients from the Original Reliability Study

COMPARISON	N = 44	COEFFICIENT OF CORRELATION
First Interview versus Second Interview (Mean of scores from both Goal Attainment Follow-up Guides from the first interview compared to means of both scores from the second interview.)		.704
Intake Interviewer Follow-up Guide, First Interview versus Second Interview		.711
Therapist Follow-up Guide, First Interview versus Second Interview		.625

When the results of this study are analyzed in terms of variance components, it is estimated that 18% of the variance is due to follow-up interviewer errors in scoring, 17% of the variance is due to the choice of material on the follow-up guide itself, 15% of the variance is due to the effects of time or circumstance difference between the two follow-up interviews, and the remaining 50% of the variance is due to client outcome.

8. The Interdisciplinary Reliability Study

This study of scorer accuracy is described more extensively elsewhere (see chapter on follow-up in P.E.P. Report 1969-1973). The study's basic goals were to compare scoring of the Goal Attainment Follow-up Guide by two methods (telephone versus in-person interviewers) and by two different types of interviewers (M.S.W. versus R.N.). In ten months, 60 clients at the Hennepin County Mental Health Service were interviewed twice. The assignment of follow-up method and type of interviewer were random. There was a mean interval of 27 days between a client's first and second interview. Table II presents the correlations obtained for the various interviewer combinations. None of the differences in mean scores reached the $p < .05$ level of significance. (Data from this study is based on Audette's chapter on follow-up procedures in the P.E.P. Report 1969-1973.)

TABLE II: Inter-Scorer Correlation Coefficients for the Interdisciplinary Study

COMPARISON	NUMBER OF CASES	CORRELATIONS	INTERVIEW	GOAL ATTAINMENT SCORE MEANS	GOAL ATTAINMENT SCORE STANDARD DEVIATION
Total, First vs Second Interview	60	.616 (sign. at $p < .001$, 2-tailed)	First	50.7	11.7
			Second	52.4	12.7
Both Interviews by M.S.W.'s, First vs Second Interview	13	.698 (sign. at $p < .02$, 2-tailed)	First	51.8	16.0
			Second	50.1	16.5
Both Interviews by R.N.'s, First vs Second Interview	10	.570 (not significant)	First	58.3	11.7
			Second	57.9	8.9
First Interview by R.N., Second Interview by M.S.W.	18	.709 (sign. at $p < .001$, 2-tailed)	First	50.0	10.0
			Second	52.2	10.0
First Interview by M.S.W., Second Interview by R.N.	19	.592 (sign. at $p < .01$, 2-tailed)	First	48.8	10.0
			Second	51.2	12.4

In the original reliability study, all interviews were by experienced social workers. In the "interdisciplinary" study, even with variation of follow-up method and type of interviewers, the reliability coefficients are similar.

C. Internal Consistency Indicators

Internal consistency is not essential to an understanding of reliability of Goal Attainment Scaling, but the internal consistency measures illustrate interesting psychometric features of the Goal Attainment Scaling methodology. Two analyses illustrating features of the Goal Attainment Scaling internal consistency are presented in this sub-section.

1. Correlations Between the Individual Scale Scores and Total Follow-up Guide Scores

The item/total correlation which is often used to suggest the degree of internal consistency can be adapted to Goal Attainment Scaling. Since there are so few items on any one Goal Attainment Follow-up Guide, one hundred Goal Attainment Follow-up Guides were selected randomly from the follow-up guides which had been scored for the Program Evaluation Project "four mode study". For these 100 follow-up guides, a correlation was calculated between the scale-by-scale score (with a range of 1 to 9) on each of the 344 scales and the overall Kiresuk-Sherman Goal Attainment score for the Goal Attainment Follow-up Guide on which that scale was contained. For example, if a follow-up guide contained three scales, one scored four, one scored five and one scored six, the following pairs would be formed: (4,50), (5, 50) and (6,50).

Such a pairing was carried out for all 344 scales in the sample, and these 344 pairs were correlated to produce the single coefficient of .693. The mean scale score was 5.34 and the mean overall, Kiresuk-Sherman Goal Attainment Follow-up Guide score was 52.19 (Data from calculations by C. Jasperson and J. Baxter). This coefficient suggests a moderately high correlation on the average between any single scale score and the corresponding score on the Goal Attainment Follow-up Guide which contains it. Thus, to know the results of scoring one scale on the Goal Attainment Follow-up Guide, allowed a fairly good prediction of the total Kiresuk-Sherman score.

2. Correlations between the Individual Scale Scores and Total Follow-up Guide Scores, by Number of Scales

A second data analysis was performed on all 634 Goal Attainment Follow-up Guides scored for the main Program Evaluation Project study. These follow-up guides include a total of 2173 scales for a mean of 3.43 scales per Goal Attainment Follow-up Guide. In this analysis, the correlation described above, (between the score on the individual scale and the overall Kiresuk-Sherman Goal Attainment score for the follow-up guide of which the scale was part) was calculated separately for Goal Attainment Follow-up Guides with differing numbers of scales, so as to remove any confusing effect of varying numbers of scales per Goal Attainment Follow-up Guide. This precaution was not applied to the preliminary scale/total analysis described above.

All the correlation coefficients as shown on Table III are statistically significant at better than the $p < .01$ level (two-tailed).

TABLE III: Correlations Between Scale Scores and Overall Goal Attainment Scores for 644 Cases, Divided by the Number of Scales per Follow-up Guide.

Number of Scales per Follow-up Guide	Number of Follow-up Guide *	Number of Scales	Mean Scale Score	Mean Goal Attainment Score	r
1	12	12	6.50	57.5	.98
2	85	170	5.34	52.0	.78
3	262	786	5.35	52.4	.68
4	180	720	5.10	50.9	.66
5	88	440	5.06	50.7	.64

* The correlations for the 4 Goal Attainment Follow-up Guides with 6 scales and the 3 Goal Attainment Follow-up Guides with 7 scales are not included because of the small size of the samples.

This analysis suggests that for follow-up guides with two to five scales, any one scale score will tend to have a moderately high correlation with the overall Goal Attainment score for that follow-up guide. Although this internal consistency indicator is applied in a modified manner, it does suggest that there is a meaningful degree of cohesion between scale scores and total follow-up guide scores.

D. Inter-rater Reliability as Measured by Rating of the Videotape Interviews

1. Six-Rater Study

Three graduate students in social work observed the videotape of an intake interview with a 22 year old female client as part of a Program Evaluation Project training program. After this observation, they each independently constructed a Goal Attainment Follow-up Guide for the client. These were the first follow-up guides constructed by them.

At a later training session, these three graduate students plus three others watched the video tape of a follow-up with that same client. Each of these six persons scored each follow-up guide.

Two of these student-constructed follow-up guides contained four scales and the third follow-up guide contained three scales, for a total of eleven scales. Thus, there were eleven scales scored six times each, except for two instances in which a scale was not scored by the raters for unknown reasons, so that there were 64 scale scores. Clearly, this is not a typical application of Goal Attainment Scaling. It suggests, however, the reliability possible with very inexperienced Goal Attainment Follow-up Guide constructors and follow-up raters.

Content agreement on the three follow-up guides is high, as Figure II illustrates. "Self-concept" appears on all three follow-up guides. Vocational problems appear in two follow-up guides, as do both marriage and sexuality problems.

FIGURE II: Scale Headings for Three Different Follow-up Guides for the Same Client

Follow-up Guides	Scale 1	Scale 2	Scale 3	Scale 4
Follow-up Guide 1	Sexual Adjustment	Marriage Counseling	Vocational Interests	Self-Concept
Follow-up Guide 2	Marriage	Job	Capacity For Close Relationships	Self-Image
Follow-up Guide 3	Self-Concept	Goal Achievement	Sexual Relationship with Husband	

As Table IV reveals, the fifteen inter-rater percentages of complete agreement on the individual scale scores range from 36 percent to 73 percent, with a mean percentage of complete agreement of 56 percent. The percentages of scale scores on which the two raters' scores are within one goal attainment level of each other ranges from 60 percent to 100 percent, with a mean percentage of 83 percent.

TABLE IV: Inter-Rater Percentage of Agreement for Six Raters Agreement Expressed in Percentages of Scales*

Rater One \ Rater Two	A	B	C	D	E
A	60% (90%) N=10				
B	60% (90%) N=10	45% (82%) N=11			
C	50% (90%) N=10	45% (91%) N=11	73% (91%) N=11		
D	44% (67%) N=9	60% (80%) N=10	50% (50%) N=10	70% (70%) N=10	
E	60% (90%) N=10	36% (91%) N=11	64% (91%) N=11	73% (100%) N=11	50% (60%) N=10

*For each pair of raters there are two percentages: the first percentage shows the percentage of scales on which there was complete agreement, and the second percentage, which is in parentheses, shows the percentage of scales in which the two raters' scores were no more than one level apart.

2. Four-Rater Study

A set of six follow-up guides was constructed by six Hennepin County Mental Health Service clinicians, shortly after the video tape for the client mentioned previously was recorded. This study reports on a session on January 19, 1973 in which one experienced Goal Attainment Scaling rater (from the Program Evaluation Project staff) and three inexperienced Goal Attainment Scaling raters scored one set of the 29 scales included in these six follow-up guides, after observing the videotape of the follow-up interview with client P.R.

As Table V shows, the mean follow-up guide score per rater varies considerably. This variation suggests that there is a meaningful difference or bias among the raters in their overall level of scoring the follow-up guides despite the high percentage of agreement on any one scale.

TABLE V: Sum of Scale Scores** for Six Follow-up Guides Based on the Videotape as Scored by Four Raters

	Follow-up Guide 1 (4 Scales)	Follow-up Guide 2 (6 Scales)	Follow-up Guide 3 (5 Scales)	Follow-up Guide 4 (5 Scales)	Follow-up Guide 5 (4 Scales)	Follow-up Guide 6 (5 Scales)	Mean Follow-up Score per Rater
Rater A	3	0	1	1	3	3.5	1.92
Rater B	-2	-7	-1.5	0	-3	1*	-2.08
Rater C	-3	-5.5	- .5	1	0	3.5	- .75
Rater D	-1	2	-1	-1	1	2	.35
Mean Score per Follow-up Guide	-.75	-2.63	- .50	.25	.25	2.5	

* Only one of the four scales was scored, for an unknown reason.

** Based on a possible range of -2 to +2 for each scale.

Table VI incorporates the data from Table V in the form of percentages of agreement, where the percentage is based on the number of raters agreeing on the most common response.

TABLE VI: Percentages of Agreement for Twenty-nine Scales, Scored by Four Raters

Follow-up Guide	Scale Number	Most Common Score	Per Cent Agreeing
Follow-up Guide 1	1	-1	50%
	2	-1	100%
	3	0	75%
	4	0	75%
Follow-up Guide 2	5	---	0%
	6	-1	50%
	7	-2	75%
	8	0	75%
	9	-1	100%
	10	---	0%
Follow-up Guide 3	11	-1	50%
	12	---	0%
	13	0	75%
	14	0	75%
	15	0	75%
Follow-up Guide 4	16	0	50%
	17	-1	100%
	18	-2	100%
	19	1	75%
	20	2	100%
Follow-up Guide 5	21	1 and 0	50%
	22	2	50%
	23	0	50%
	24	-1 and -2	50%
Follow-up Guide 6	25	-.5	50%
	26	2	50%
	27	0	75%
	28	1	75%
	29	1	100%

The mean of these twenty-nine percentages of agreement is 66.7. There was complete agreement on six scales, 75 percent agreement on ten scales, 50 percent agreement on ten scales and 0 agreement (i.e., all four ratings were different) on three scales.

However, even with this moderately high percentage of agreement, the sum of all the scale scores varied greatly among the four raters.

E. Inter-rater Reliability in Two Goal Attainment Follow-up Guides Constructed to Evaluate Organizational Goal Achievement

The idea of applying Goal Attainment Scaling to organizations is an addition to the original concept of evaluating individual clients. However, the concepts of organizational goal-setting and individual goal-setting are basically similar. The inter-rater reliability of two such organizational applications are presented below.

1. A Twenty-three Scale Goal Attainment Follow-Up Guide

A Goal Attainment Follow-up Guide was constructed for the Program Evaluation Project in 1971 by the supervisory staff. There were 23 scales utilized in follow-up scoring.

This twenty-three scale Goal Attainment Follow-up Guide was scored by eleven persons from all levels of the Program Evaluation Project staff. Not every scale was scored by all eleven because raters were given the option of not registering a rating if they considered the scale to be unscoreable.

The mean percentage of agreement for these scales is 66.9%, which is very similar to the earlier mean percentage of agreement 66.7% calculated for 29 scales in the four-rater study reported in section D-2.

2. A Sixty Scale Goal Attainment Follow-up Guide

A set of 60 scales was constructed by the Program Evaluation Project supervisory staff for a 1972 evaluation of the organization's goal attainment. The resulting Goal Attainment Follow-up Guide was scored by seven Program Evaluation Project staff members and, as in the earlier study, they were given the option of not rating a scale. For some scales, there was high agreement that the scales were unscoreable, and where four or more staff members concur in rating a scale "unscoreable" the number of "unscoreable" ratings is used to calculate the percentage of agreement.

The mean percentage of agreement for these 60 scales was 68.1, which is similar to the other percentages of agreement reported above.

F. Reliability of Scores When Multiple Follow-Up Guides are Constructed for the Same Client

The original reliability study was based on a design in which two Goal Attainment Follow-up Guides were made for each client. One follow-up guide was constructed by the therapist and one was constructed by the intake interviewer. There was a mean of 25 days between construction of the two follow-up guides. The two follow-up guides were developed in two different settings (i.e., intake versus therapy interviews). In the studies reported below, which are based on the videotape, several follow-up guides could be based on the single, recorded intake interview, thus minimizing or eliminating effects of different times and settings. However, since the follow-up guide constructors were inexperienced and the videotape interview is obviously a different situation than a live interview, the results are not directly comparable.

1. The Six-Rater Study

In this study, as presented below, (see section D-1) three Goal Attainment Follow-up Guides were constructed after the raters observed the videotape on client P.R. The follow-up guide constructors were inexperienced, both as constructors and clinicians. Similarly, the six persons who scored these three Goal Attainment Follow-up Guides had not previously rated any follow-up guides of this type. Thus, these results reflect the reliability of follow-up guide construction and of follow-up guide scoring without the benefits of experience except for minimal training in Goal Attainment Scaling (approximately one and one half hours training in follow-up guide scoring).

For each of these independently constructed follow-up guides, the mean score per scale was calculated and appears on Table VII below.

TABLE VII: Mean per Scale Ratings* for Three Follow-up Guides Based on the Videotape and Scored by Six Raters

Rater	Guide 1	Guide 2	Guide 3
A	-1.33	0	- .33
B	- .25	.75	0
C	-1.00	.25	- .67
D	-1.00	.75	- .67
E	.50	1.33	0
F	-1.00	.75	- .67

* Based on a -2 to +2 possible range for a single scale.

The correlations among the six ratings for each follow-up guide appear on Table VIII below. These are the intercorrelations (Pearson Product Moment) of the columns in Table VII.

TABLE VIII: Correlations Among the Mean Per Scale Scores for Six Independent Scores on Three Follow-up Guides

COLUMNS CORRELATED	COEFFICIENT
Follow-up Guide 1 and Follow-up Guide 2	.760
Follow-up Guide 1 and Follow-up Guide 3	.765
Follow-up Guide 2 and Follow-up Guide 3	-.234
Mean of the Absolute Value of the Three Correlations	.586

The mean absolute value of the correlation coefficient is .586 which suggests the stability of the mean scale score per follow-up guide construction even for the very inexperienced constructors in this example involving a videotape case study of one client. The range of correlations among the pairs of raters, from -.234 to .765 suggests that individual differences in agreement among the follow-up guide constructors are quite high. The .760 correlation for follow-up guide constructor 1 versus follow-up guide constructor 2, and .765 correlation for follow-up guide constructor 1 versus follow-up guide constructor 3 are comparable to the correlation coefficients calculated for data of the original reliability study. (See Table I.)

G. Reliability of Different Follow-up Raters

1. Client Scoring versus Follow-Up Interviewer Scoring for Client-Constructed Follow-Up Guides (Guide to Goals Study)

The Guide to Goals, Format One, can be used to enable clients to construct their own follow-up guides. In the Day Treatment Center of the Hennepin County Mental Health Service, clients were asked to use the Guide to Goals.

These client-produced follow-up guides were scored independently by 1) the client and 2) by the follow-up interviewer at an interview scheduled four months after the date at which the follow-up guide was constructed. The interviewers were from the regular Program Evaluation Project follow-up staff. Goal Attainment scores, one from the client and the other from the therapist, were correlated. The Guide to Goals study was divided into two phases, and correlations from the cases first followed-up were published in the Program Evaluation Project Newsletter in Volume IV, issues 4 and 6 (Jones and Garwick).

In the report for phase one, scores from ten cases were available. The clients' ratings and the interviewers' ratings were correlated (Pearson Product Moment) with a coefficient of .71, which is statistically significant at the $p < .01$ level, two-tailed.

In phase two, only half of the clients were invited to construct their own follow-up guides. Interviewer and client-rated Goal Attainment scores for seven client-constructed follow-up guides were correlated at .733 which is statistically significant at the $p<.03$ level, two-tailed. The means were 71.6 for the client-rated Goal Attainment score and 69.9 for the interviewer-rated Goal Attainment score.

In these two studies, both ratings took place at the same session. The inter-rater reliability is comparable to the reliability coefficient obtained from other studies, even though the follow-up guides were prepared by persons so incapacitated that they sought assistance at the Mental Health Day Treatment Center, and who were constructing their first follow-up guide.

2. Therapist Scoring Versus Follow-up Worker Scoring

This study involved the therapists' scoring of follow-up guides which had been constructed at the Hennepin County Mental Health Service by intake interviewers. (Baxter, 1973) After a follow-up interviewer had scored the follow-up guide and returned it to the Program Evaluation Project staff, the follow-up guide was given to the therapist for scoring, before the therapist saw the follow-up interviewer's scoring. The therapist did not interview the client when scoring the Goal Attainment Follow-up Guide. In practice, there was a sizeable delay of at least one or two weeks between these two ratings, although the actual length of time is not specified in the original report. Under these conditions, the correlations between therapist and follow-up interviewer ratings tend to be quite modest for the Adult Outpatients and quite high for the Day Treatment clients. (See Table IX.)

TABLE IX: Pearson Correlation Between Goal Attainment Scores for Clients Based on Two Independent Interviews, One by Therapist and One by Follow-up Interviewer

Type of Client	N	Correlation
Adult Outpatient Clients	N = 53	.507
Day Treatment Clients	N = 8	.848

H. Reliability Over Time (The Drug Effectiveness Study)

The Drug Study procedures are described in another chapter of the P.E.P. Report 1969-1973. However, the study involved three follow-up interviews per client. One interview was held three weeks after construction of the follow-up guide, the second was two months after construction, and the third was six months after construction. The follow-up guides were focused most directly on the two month follow-up interview. Follow-up interviews were conducted by experienced masters degree social workers on the Program Evaluation Project follow-up staff.

As the data in Table X on the following page suggest, there appears to be a trend for the mean Goal Attainment scores to rise as the time after follow-up increases. However, the rise in means would not necessarily be related to changes in reliability, if all cases increased proportionately. Table X shows the correlations of scores for the follow-up dates.

TABLE X: Mean Goal Attainment Scores for Three Follow-up Sessions

Follow-up Session	Mean (Kiresuk-Sherman) Goal Attainment Score	Standard Deviation	N
Three Weeks After Follow-up Guide Construction	51.17	10.8	18
Two Months After Follow-up Guide Construction	56.28	11.0	17
Six Months After Follow-up Guide Construction	61.3	10.1	16

As the correlations in Table XI below reveal, there is a relatively low degree of correlation among the Goal Attainment scores from the three follow-up times. The mean correlation is only .46, which is lower than correlations typically obtained from multiple follow-ups at the same follow-up session or from two follow-ups by different follow-up workers. This somewhat lower correlation may imply that the degree of attainment of expectations does change considerably within a six month period, that is, that knowledge of goal attainment at one time is not always sufficient for prediction of goal attainment at other times. It appears that in addition to the rise in mean Goal Attainment scores over time, there is a fairly high amount of change in the relative degree of goal attainment over time among the clients.

TABLE XI: Correlations of Kiresuk-Sherman Goal Attainment Scores for Three Follow-up Interviews

	Follow-up Two Months After Follow-up Guide Construction	Follow-up Six Months After Follow-up Guide Construction
Follow-up Three Weeks After Follow-up Guide Construction	.70 (N=17) (p<.001, two-tailed)	.20 (N=14) (not statistically significant at the p < .05 level)
Follow-up Two Months After Follow-up Guide Construction		.47 (N=16) (p<.05, two-tailed)

III. An Overview of Correlational Estimates of Goal Attainment Scaling Reliability

As emphasized in the introductory comments, the complexity of reliability estimation should now be clearer. Goal Attainment Scaling has many points at which the reliability of the technique can be altered. These points include, at least:

a. who constructs the follow-up guide? (client, therapist or intake interviewer, etc.)

b. how many scales appear on the follow-up guide?

c. when is the follow-up guide scored? (how long after follow-up guide construction?)

d. who scores the follow-up guide? (are they experienced in such scoring? what background do they have?)

e. how many raters are there?

f. in what circumstances are the follow-up guides scored? (in-person, by mail, by telephone)

Since Goal Attainment Scaling is a basic approach which encompasses an array of possible technical variants, a whole galaxy of possible reliability estimates could be helpful for different applications.

The reliability estimates discussed in this chapter were produced in a variety of situations but most suggest a usefully high degree of inter-rater agreement, and a usefully high degree of Goal Attainment score stability for clients when follow-up guides are made by more than one person.

References.

Anastasi, A. Psychological testing. USA: Macmillan Co., 1968, Pp. 71, 81, 94.

Baxter, J. Combination validity/reliability study. Unpublished Program Evaluation Project report, February, 1973.

Cronbach, L. J. Coefficient alpha and the internal structure of tests. Psychometricka, 1951, 297-334.

Dahlstrom, W. G. Recurrent issues in the development of the MMPI. In J. N. Butcher (Ed.), MMPI: research developments and clinical application. USA: McGraw-Hill, 1969, 27-28.

Hathaway, S. R. Scales 5 (masculinity-femininity), 6 (paranoia) and 8 (schizophrenia). In G. W. Welsh and W. G. Dahlstrom (Eds.), Basic readings on the MMPI in psychology and medicine. Minneapolis, Minnesota: University of Minnesota Press, 1956, 11.

Holzberg, J.D. Reliability re-examined. In M. S. Rickers-Ovsiankina (Ed.), Rorschach psychology. New York: John Wiley and Sons, 1960, 368, 377.

Kelly, E. L. Assessment of human characteristics. Belmont, California: Brooks (Cole Publishing Company), 1967, 26.

Nunnally, J. C. Psychometric theory. USA: Mc-Graw-Hill, 1967, 20.

Sherman, R., Baxter, J. & Audette, D. Analysis of the original reliability study of goal attainment scaling. P.E.P. Report 1969-1973, Chapter Four.

PROGRAM EVALUATION PROJECT STAFF LISTING

CURRENT STAFF MEMBERS

Thomas J. Kiresuk, Ph.D.
Principal Investigator 1969-1974

Donna M. Audette
Research Assistant 1970
Follow-up Assistant 1971
Follow-up Supervisor 1972-1973
Utilization Consultant 1974

James Baxter
Research Assistant 1970
Operations Manager 1971
Assistant Coordinator 1972-1973
Re-design Coordinator 1974

Diane Berg
Editorial Secretary 1973-1974

Janis Bibee, M.A.
Editorial Secretary 1972
Editorial Assistant 1973
Assistant to the Editor 1974

Joan Brintnall
Student Assistant 1971
Secretary-Receptionist 1972
Administrative Assistant 1973
Dissemination, Consultation,
and Utilization Supervisor 1974

Joan Dreyer
Research Clerk 1972
Research Assistant 1973
Research/Administrative Clerk 1974

Geoffrey Garwick, M.A.
Research Applications Consultant 1970
Program Evaluation Coordinator 1971
Deputy Assistant Director 1972
Assistant Director 1973
Dissemination, Consultation,
and Utilization Consultant 1974

Carolyn Jasperson
Student Assistant 1973-1974

Susan Jones
Research Assistant 1971
Research Associate 1972-1974

Laurence Kivens, M.A.
Applications Analyst 1971
Editorial Supervisor 1971
Editor 1972-1974

Mary Knepper
Appointment Interviewer 1970-1973
Administrative Assistant 1974

Judy Long
Administrative Secretary 1974

Sander Lund
Management Applications Supervisor 1971
Assistant Coordinator 1971-1972
Coordinator for Administration 1973
Assistant Director 1974

Nancy Petersen
Secretary 1973
Research Assistant 1974

Michael Saunders
Programmer Analyst 1971-1972
Programmer Supervisor 1973-1974

Robert Sherman, Ph.D.
Associate Investigator 1969-1974

Roger Twedt
Student Assistant 1974

Mary Ellen Whalen
Student Assistant 1973-1974

PREVIOUS STAFF MEMBERS

Anita Bjornson
Research Assistant 1970

Barbara Blazick
Student Assistant 1970

Mary Duroche
Editorial Secretary 1972

David Feigal
Research Assistant 1970-1971

Thomas Griffin
Student Assistant 1970

Marilee Grygelko
Student Assistant 1971-1972
Research Assistant 1973

Edward Gubman
Student Assistant 1971

Colleen Halley
Student Assistant 1971-1973

Robert Kearney
Editorial Assistant 1972

Karen Kohout, M.A.
Research Analyst 1969
Research Supervisor 1970-1971

Sherry Lampman
Administrative Assistant 1970
Linguistic Analysis Consultant 1971
Content Analysis Supervisor 1972

William Makela, M.A.
Follow-up Supervisor 1970-1972

Charles Meade
Research Assistant 1970-1972

Deirdre Meade
Secretary-Receptionist 1971
Administrative Assistant 1972

Sylvia Muilenberg
Administrative Assistant 1969
Administrative Supervisor 1970

Nils Olsson
Student Assistant 1971
Research Assistant 1972-1973

Carol Pollock
Research Clerk 1972

William Prock
Community Applications Supervisor 1971-1972

Peter Ree
Student Assistant 1970

Martha Rosen
Secretary-Receptionist 1970
Editorial Assistant 1971-1972

Susan Salasin
Coordinator 1969-1970
Assistant Director 1971-1972
Research Applications Consultant 1973

Richard Tripp
Programming Supervisor 1970
Design and Analysis Supervisor 1971-1972

Mary Trone
Editorial Secretary 1973

Carol Vanderpool
Secretary-Receptionist 1971-1972
Administrative Assistant 1973

Cynthia Wetterland
Secretary-Receptionist 1970

Allen Wichelman
Medical Records Clerk 1970
Management Applications Supervisor 1971

Sue Wright
Clinical Applications 1971

Carole Zimbrolt
Research Analyst 1969-1971

FOLLOW-UP INTERVIEWERS - CURRENT

Kathleen Bergum, M.S.W.

Charles Besnett, M.S.W.

Carol Dethmers, B.A.

Marcia Frankenberg, B.A.

George Meirick, M.S.W.

FOLLOW-UP INTERVIEWERS PREVIOUS

Mary Ann Anzelc, R.N.

James Bergum, M.S.W.

Roanne Borkon, R.N.

Larry Bultena, M.S.W.

Scott Craven

Jeanine Emmons, R.N.

Barbara Gusek, R.N.

Mary Keturakat, R.N.

Steve Lapinsky

Betty Metz, B.A.

Madiline Sachs, R.N.

James Snope, M.S.W.

Milt Somerfleck, M.S.W.

CONSULTANTS

Dean Beaulieu, Ph.D.
Clinical Coordinator 1971-1974

James Boen, Ph.D.
Statistical Consultant 1969-1973

Byron Brown, Ph.D.
Statistical Consultant 1969-1973

Arthur Funke, Ph.D.
Dissemination Consultant 1969-1974

Stephen Greenwald, M.D.
Medications Consultant 1969-1973

Ann Russell
Clinical Consultant 1970

Robert Spano, A.C.S.W.
Patient Follow-up Coordinator 1969-1973

Wyman Spano
Editorial Consultant 1972-1973

Robert Walker, M.A.
Data Applications Coordinator 1973